

OPTIMALITY REGIONS AND FLUCTUATIONS FOR BERNOULLI LAST PASSAGE MODELS

NICOS GEORGIOU AND JANOSCH ORTMANN

ABSTRACT. We study the sequence alignment problem and its independent version, the discrete Hammersley process with an exploration penalty. We obtain rigorous upper bounds for the number of optimality regions in both models near the soft edge. At zero penalty the independent model becomes an exactly solvable model and we identify cases for which the law of the last passage time converges to a Tracy-Widom law.

1. INTRODUCTION

1.1. Directed growth models. In this article we study a generalisation of two specific models of directed last passage percolation, namely the *longest common subsequence model* concerning the size of the longest common subsequence between two uniformly drawn words from a finite alphabet [8], and an independent version, which is exactly solvable. We call the latter the *independent model* and it is a discrete analogue of the Hammersley process [20], introduced in [40]. We study these models near directions for which the corresponding shape function starts developing a flat segment, which is called the *soft edge* of the model. Both models fit in the general framework [14], namely there is:

- (i) The random environment $\omega \in \mathbb{R}^{\mathbb{Z}^2}$, whose law we denote by \mathbb{P} . Each marginal ω_u should be viewed as a random weight placed on site $u \in \mathbb{Z}^2$.
- (ii) A collection Π of admissible paths on \mathbb{Z}^2 . A path π from u to v is uniquely identified by an ordered sequence of integer sites, so when necessary we write $\pi = \{u = u_0, u_1, \dots, u_\ell = v\}$. A path π is admissible if and only if its increments $z_k = u_k - u_{k-1}$ are contained in a finite set $\mathcal{R} \subset \mathbb{Z}^2$. For $u, v \in \mathbb{Z}^2$ we denote the set of admissible paths from u to v by $\Pi_{u,v}$. It is a requirement that \mathbb{P} is stationary and ergodic under shifts $T_z, z \in \mathcal{R}$.
- (iii) A measurable potential function $V : \mathbb{R}^{\mathbb{Z}^2} \times \mathcal{R}^\ell \rightarrow \mathbb{R}$. For the two models under investigation we always have $\ell = 1$ and V is a bounded function, thus satisfying the technical assumptions of [14].

The *point-to-point last passage time* from u to v is the random variable G^V defined by

$$(1.1) \quad G_{u,v}^V = \max_{\pi \in \Pi_{u,v}} \left\{ \sum_{u_k \in \pi} V(T_{u_k} \omega, z_{k+1}) \right\}.$$

2000 *Mathematics Subject Classification.* 60K35.

Key words and phrases. Soft edge, edge results, optimality regions, sequence alignment, discrete Hammersley process, longest common subsequence, Bernoulli increasing paths, Tracy-Widom distribution, last passage time, corner growth models.

NG was partially supported by the University of Sussex Strategic development Fund (SDF). JO was partially supported by an ISM-CRM fellowship.

A well studied version of the model is the *corner growth model*, for which $\mathcal{R} = \{e_1, e_2\}$, the coordinates of ω are i.i.d. under \mathbb{P} and $V(\omega, z) = \omega_0$. It is expected that under some regularity assumptions on the moments and continuity of ω_0 , the asymptotic behaviour of G^V (e.g. fluctuation exponents for G^V and the maximal path, distributional limits, etc) is environment-independent. This is suggested by results available for the two much-studied exactly solvable models when ω_0 is exponentially or geometrically distributed and further evidenced by the general theory in [14–16] and the edge results of [7, 31], as we discuss later.

In this article, the set of admissible steps is $\mathcal{R} = \{e_1, e_2, e_1 + e_2\}$ and the coordinates of the environment take values in $\{0, 1\}$. Our choice of potential is a two-parameter family of bounded functions, indexed by two non-negative parameters α and β :

$$(1.2) \quad V_{\alpha, \beta}(\omega, z) = \begin{cases} \omega_0 - \alpha(1 - \omega_0) & \text{if } z = e_1 + e_2 \\ -\beta & \text{if } z \in \{e_1, e_2\}. \end{cases}$$

This particular choice of potential is inspired by a problem which appears in computational molecular biology, computer science and algebraic statistics, as we now explain.

1.2. The alignment model. The problem of *sequence alignment* [34, 42] can be cast in this framework. Consider two words $\eta^x = \eta_1^x \dots \eta_m^x$ and $\eta^y = \eta_1^y \dots \eta_n^y$ formed from a finite alphabet \mathcal{A} . We are looking for a sequence of elementary operations of minimal cost that transform η^x to η^y . These operations are:

- (1) replace one letter of η^x by another, at a cost α
- (2) delete a letter of η^x or insert another letter, each at a cost of β .

Assign a score of 1 for each match and subtract the costs for replacements, deletions and insertions. Each sequence of operations taking η^x to η^y is thus assigned a *score* $L_{m,n}^{(\alpha, \beta)}$, also often called the *objective function*. We will also write $L_{m,n}^{(\beta)}$ for $L_{m,n}^{(0, \beta)}$.

A problem arising in molecular biology [1, 21, 35, 37, 44, 45] is to maximise this alignment score. In that context the words η^x and η^y can be DNA strands (with $\mathcal{A} = \{A, C, G, T\}$), RNA strands ($\mathcal{A} = \{A, C, G, U\}$) or proteins (with \mathcal{A} the set of amino acids that make up a protein), and the elementary operations correspond to mutations. A choice of the parameters α and β corresponds to a judgement on how frequently each type of mutation occurs. The optimal score for an alignment of η^x with η^y can then be considered a measure of similarity between these words. The question also appears in algebraic statistics [38]: there the objective function is the tropicalisation of a co-ordinate polynomial of a particular hidden Markov model.

The special case $\alpha = \beta = 0$ corresponds to the problem of finding longest common subsequence (LCS) of the words η^x and η^y , which has been intensively studied by computer scientists [6, 22, 29, 32] and mathematicians [2, 8, 18, 23, 27, 28].

On the other hand, the alignment score $L_{m,n}^{(\alpha, \beta)}$ is the last passage time (1.2) in environment

$$(1.3) \quad \omega_{ij} = \begin{cases} 1 & \text{if } \eta_i^x = \eta_j^y \\ 0 & \text{otherwise,} \end{cases}$$

i.e. the marginals of ω are (correlated) Bernoulli random variables with parameter $|\mathcal{A}|^{-1}$. The model with this choice of environment is referred to as the *alignment model*.

A deletion of a character in η^x corresponds to a horizontal step (e_1) in the last passage model, whereas an insertion of a letter into η^x corresponds to a vertical step (e_2). Replacing a letter in η^x by another corresponds to a diagonal step ($e_1 + e_2$) onto a point (i, j) where $\omega_{ij} = 0$, whereas any letter left alone (i.e. a successful alignment) corresponds to a diagonal step onto a point (i, j) where $\omega_{ij} = 1$. The path in Figure 1 corresponds to the alignment

$$\begin{aligned}\eta^x &: A \bar{A} B A - B A \\ \eta^y &: - A B A \bar{A} B A\end{aligned}$$

in which the bar under the first A of η^x corresponds to deleting the letter A from η^x while the bar in η^x corresponds to inserting the letter A there. A convenient way to look at this is that the bars, called *gaps*, are used to stretch the two words appropriately so that different matchings are obtained.

Which paths are optimal depends on the choice of parameters α, β . In molecular biology these parameters are often chosen ad hoc and it is not clear that there is a single ‘right’ choice [44]. An alternative approach is to consider the space $\mathcal{C} = [0, \infty) \times [0, \infty)$ of all possible parameters (α, β) and to analyse how the optimal paths change as (α, β) varies. A maximal subset of \mathcal{C} on which the set of optimal paths does not change is called an *optimality region* of \mathcal{C} . The shape of optimality regions in \mathcal{C} are semi-infinite cones bounded by the coordinate axes and by lines of the form $\beta = c + \alpha(c + 1/2)$ for certain values of c . So it suffices to study the number of regions with one parameter fixed; we will set $\alpha = 0$.

Denote the number of optimality regions in this model by $R_{m,n}^{(\text{al})}$. Naturally the (expected) number of optimality regions attracted a lot of interest both theoretically [12, 19] and in applications [10, 24, 30, 33]. The current conjecture [11, 38] is that $\mathbb{E}(R_{n,n}) = O(\sqrt{n})$, but the complexity of the random variable does not allow for calculations.

We consider the case where each letter of η^x and η^y is chosen independently and uniformly at random from \mathcal{A} . The lengths of η^x and η^y are n and $|\mathcal{A}|n - xn^a$ respectively, where a is in the interval $(0, 1)$. We obtain an asymptotic lower bound for the optimal score when α is fixed, as well as upper bounds for the number of optimality regions.

1.3. The independent model. In the alignment model described above, the environment consists of heavily correlated random variables. For example if $i < j$ and $k < \ell$ then $\{\omega_{ik}, \omega_{il}, \omega_{jk}, \omega_{j\ell}\}$ are not independent: in fact

$$(1.4) \quad \mathbb{P}\{\omega_{ik} = 1 \mid \omega_{jk} = \omega_{il} = \omega_{j\ell} = 1\} = 1,$$

see for example in Figure 1 the weights on $(1, 1), (2, 1), (1, 3), (2, 3)$.

Stronger results can be obtained if the marginals of ω are i.i.d. Bernoulli random variables on $\{0, 1\}$ with parameter $p \in (0, 1)$, because we then obtain a solvable model [39]. This will be referred to as the *independent model*, and the score from $(0, 0)$ to (m, n) is denoted $G_{m,n}^{(\alpha, \beta)}$, with $G_{m,n}^{(\beta)} = G_{m,n}^{(0, \beta)}$.

The model was first introduced by Seppäläinen [40] as a discretisation of the Hammersley process and further studied for the case $\alpha = \beta = 0$ in [5, 13]. Asymptotic results as p tends to zero were obtained in [25].

In this paper we consider a rectangle of height n and width $m_n = n/p - xn^a$ for $a \in (0, 1)$. We obtain a tightness result for the last passage path and upper bounds for the number of

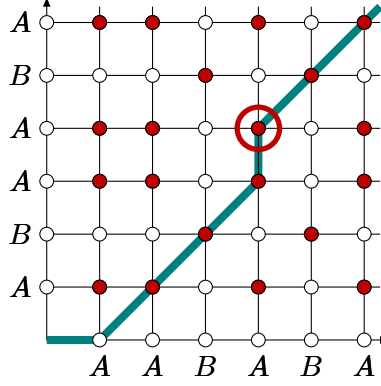


FIGURE 1. Environment generated by the two strings $AABABA$ and $ABAABA$. Coloured dots correspond to the value 1, white dots to the value 0. The thickset path is a maximal path in this environment, from $(0,0)$ to $(6,6)$ with minimal number of vertical or horizontal steps (just 2 in this case). The path collects only 5 positive weights, since the step that leads from $(4,3)$ to $(4,4)$ (the site in the thick-set circle) is vertical, while the environment only contributes to the weights if collected by a diagonal step. When $\alpha = 0$ the illustrated path has weight $5 - 2\beta$ and coincides with the last passage time for $\beta \leq 1/2$. For $\alpha = 0$ and $\beta > 1/2$ the main diagonal is optimal, with last passage time equal to 4. These are the only two optimal paths, so there are two optimality regions.

optimality regions $R_{m,n,n}^{(\text{ind})}$. We also show that the fluctuations of $R_{m,n,n}^{(\text{ind})}$ converge, suitably rescaled, to the Tracy–Widom GUE distribution.

1.4. Edge results. There is a collection of results in the corner growth model literature that are classified as *edge results*. The terminology is motivated by the fact that the last passage time T is studied in a thin rectangle, either with dimensions $n \times yn$ and letting $y \rightarrow 0$ after sending $n \rightarrow \infty$ [31], or with only one macroscopic edge, namely of dimensions $n \times xn^\gamma$ with $\gamma < 1$. Several results near the edge are universal, in the sense that they do not depend on the particular distribution of the environment. In the sequence we denote the environment for the corner growth model by $\zeta = \{\zeta_u\}_{u \in \mathbb{Z}_+^2}$. An approximation of i.i.d. sums with a Brownian motion [26] was used in [17] to obtain the weak law of large numbers,

$$\frac{T_{n,xn^\gamma} - n\mathbb{E}(\zeta_0)}{\sqrt{\text{Var}(\zeta_0)n^{1+\gamma}}} \Rightarrow c\sqrt{x}, \quad (n \rightarrow \infty),$$

and simulations lead to the conjecture that $c = 2$. The conjecture was proved in [41] via a coupling with an exclusion process and later in [4] using a random matrix approach. A coupling with the Brownian last passage percolation model [4, 36] allow [7] to obtain

$$(1.5) \quad \frac{T_{n,n^\gamma} - n\mathbb{E}(\zeta_0) - \sqrt{\text{Var}(\zeta_0)n^{1+\gamma}}}{n^{\frac{1}{2}-\frac{\gamma}{6}}\sqrt{\text{Var}(\zeta_0)}} \Rightarrow W, \quad (n \rightarrow \infty),$$

where W has the *Tracy-Widom GUE distribution* [43]: the limiting distribution of the largest eigenvalue of a GUE random matrix. If ζ_0 has exponential moments, (1.5) holds for all $a \in (0, 3/\gamma)$.

There is a coupling of $G_{\frac{n}{p}-xn^a,n}^{(0)}$ with $T_{n,n^{2a-1}}$, which we describe in Section 5. This mapping was exploited in [13] to obtain the local weak law of large numbers

$$(1.6) \quad \lim_{n \rightarrow \infty} \mathbb{P} \left\{ \left| \frac{n - G_{\frac{n}{p}-xn^a,n}^{(0)}}{n^{2a-1}} - \frac{(px)^2}{4(1-p)} \right| < \varepsilon \right\} = 1$$

for all $a \in (1/2, 1)$. We use the same coupling to obtain a distributional limit for the edge.

1.5. Outline. The paper is organised as follows: in Section 2 we state our main results. Section 3 contains preliminary results that do not depend on the specific choice of environment and therefore hold for both the alignment and the independent model. In Section 4 we prove our results about the alignment model whereas the results concerning the independent model are proved in Sections 5 and 6.

2. RESULTS

In this section we state our main results, first for the alignment, then for the independent model. Throughout we fix a finite alphabet \mathcal{A} with $|\mathcal{A}| \geq 2$, from which the letters of words η^x and η^y are chosen uniformly at random, independently of each other and let $a \in (0, 1)$ and $\alpha, \beta \geq 0$. The proofs of Theorems 2.1, 2.2 and 2.3 can be found in Section 4. See Section 5 for a proof of Theorems 2.5, 2.6 and 2.7 and Section 6 for the proof of Theorem 2.8.

2.1. Alignment model. Our first result concerns the passage time in the alignment model.

THEOREM 2.1. *Let $x > 0$. All constants below depend on x , \mathcal{A} and a .*

(1) *If $a \leq 1/2$ then for any $x > 0$ and any $\beta \geq 0$ there exists a constant C_1 such that*

$$(2.1) \quad \overline{\lim}_{n \rightarrow \infty} \frac{n(1 + \beta - |\mathcal{A}|\beta) - L_{|\mathcal{A}|n-xn^a,n}^{(\beta)}}{\sqrt{n \log n}} < C_1, \quad \mathbb{P} - a.s.$$

(2) *If $a > 1/2$ then for any $x > 0$ and any $\beta \geq 0$ there exists a constant C_2 such that*

$$(2.2) \quad \overline{\lim}_{n \rightarrow \infty} \frac{n(1 + \beta - |\mathcal{A}|\beta) - L_{|\mathcal{A}|n-xn^a,n}^{(\beta)}}{n^a} < C_2, \quad \mathbb{P} - a.s.$$

Next we turn to the number of optimality regions for the alignment model. Our first result gives an upper bound on the asymptotic growth of the number regions:

THEOREM 2.2. *Let $x > 0$. All constants below depend on x , \mathcal{A} and a .*

(1) *If $a \leq 1/2$ then for any $x > 0$ and any $\beta \geq 0$ there exists a constant C_3 such that*

$$(2.3) \quad \overline{\lim}_{n \rightarrow \infty} \frac{R_{|\mathcal{A}|n-xn^a,n}^{(\text{al})}}{(n \log n)^{1/3}} < C_3, \quad \mathbb{P} - a.s.$$

(2) *If $a > 1/2$ then for any $x > 0$ and any $\beta \geq 0$ there exists a constant C_4 such that*

$$(2.4) \quad \overline{\lim}_{n \rightarrow \infty} \frac{R_{|\mathcal{A}|n-xn^a,n}^{(\text{al})}}{n^{2a/3}} < C_4, \quad \mathbb{P} - a.s.$$

We also have a bound for the expected number of optimality regions:

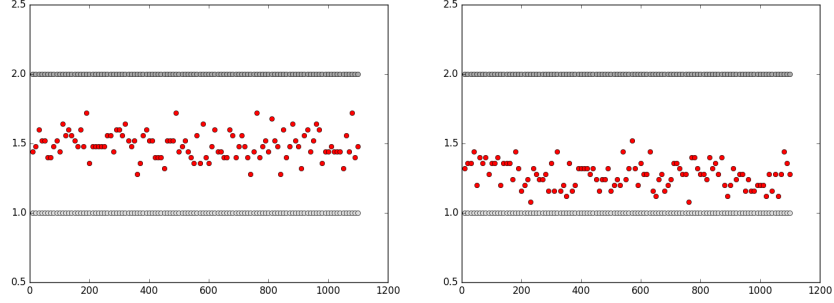


FIGURE 2. Monte Carlo simulations for the empirical maximum, minimum and expected number of regions for up to $n = 1000$ in the alignment model for small values of a . For each n , 25 independent pairs of strings were uniformly chosen. n grows in increments of size 10
 (Left) $|\mathcal{A}| = 20, a = 1/3, x = 1$. (Right) $|\mathcal{A}| = 2, a = 1/2, x = 1$. The simulations suggest the expected number of regions is bounded, and in agreement with the theoretical bound obtained for the independent model.

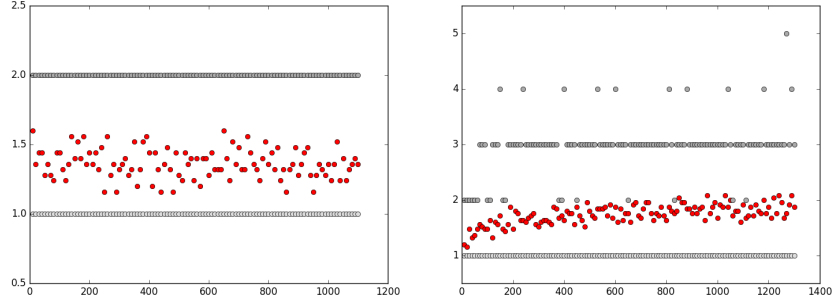


FIGURE 3. Monte Carlo simulations for the empirical maximum, minimum and expected number of regions in the alignment model when a is close to 1. For each n , 25 independent pairs of strings were uniformly chosen. n grows in increments of size 10.
 (Left) $|\mathcal{A}| = 20, a = 0.8, x = 1$. (Right) $|\mathcal{A}| = 2, a = 0.8, x = 1$. The simulations suggest that the expected number of regions is bounded for large alphabet sizes, but for small size alphabets we see growth.

THEOREM 2.3. *Let $a \in (0, 1)$. For any $x > 0$ there exists a deterministic constant $C = C(|\mathcal{A}|, x, a)$ so that*

$$(2.5) \quad \mathbb{E} \left[R_{|\mathcal{A}|n-xn^a, n}^{(\text{al})} \right] \leq \begin{cases} C\sqrt{n \log(\log n)} & \text{for } a \leq 1/2 \\ Cn^a & \text{for } a > 1/2. \end{cases}$$

Remark 2.4. These results are also valid for the independent model. Given the stronger bounds for the independent model, we do not expect (2.5) to be sharp, particularly for small values of the exponent a , and this is supported by Monte Carlo simulations. For example these suggest that for $a \leq 1/2$ the number of expected regions is bounded (see

Figure 2). This is also the case for the independent model as we see in Theorem 2.6. For $a > 1/2$, the simulations in Figure 3 show that the expected number of regions is growing for small alphabet sizes, but again the exponent of growth is smaller than $2a/3$ and it seems to depend on the alphabet size.

2.2. Independent model. We now turn to the model with independent weights. We consider the last passage time $G_{m,n}^{(0)}$ with $m_n = n/p - xn^a$ for suitably chosen x . When the exponent a is small we obtain tightness without rescaling, for any choice of x .

THEOREM 2.5. *Let $x \in \mathbb{R}$ and $a \in (0, \frac{1}{2})$. The sequence $(n - G_{n/p - xn^a, n}^{(0)})_{n \in \mathbb{N}}$ is tight and*

$$(2.6) \quad \overline{\lim}_{n \rightarrow \infty} \mathbb{P} \left\{ n - G_{n/p - xn^a, n}^{(0)} \geq k \right\} \leq \begin{cases} 2^{-k}, & a < 1/2 \\ (\Phi(xp^{5/2}q^{-2}))^k, & a = 1/2. \end{cases}$$

We will see in (3.12) that $R_{m,n}^{(\text{ind})} < n - G_{m,n}^{(0)}$. As a corollary we obtain an asymptotic bound on the expected number of optimality regions:

THEOREM 2.6. *Let $a \in (0, 1/2]$. Then*

$$(2.7) \quad \overline{\lim}_{n \rightarrow \infty} \mathbb{E} \left[R_{n/p - xn^a, n}^{(\text{ind})} \right] \leq \begin{cases} 2, & a < 1/2 \\ (1 - \Phi(xp^{5/2}q^{-2}))^{-1}, & a = 1/2. \end{cases}$$

For $a > 1/2$ we state a bound on the number $R_{m,n}^{(\text{ind})}$ of optimality regions. The optimal results and the relevant scaling of m in terms of n differ according to the value of a .

THEOREM 2.7. *Let $a \in (0, 1)$.*

(1) *If $a \in (0, 1/2]$,*

$$(2.8) \quad \overline{\lim}_{n \rightarrow \infty} \mathbb{P} \left\{ R_{p^{-1}n - xn^a, n}^{(\text{ind})} \geq k \right\} \leq \begin{cases} 2^{-k}, & a < 1/2 \\ (\Phi(xp^{5/2}q^{-2}))^k, & a = 1/2. \end{cases}$$

(2) *If $a \in (1/2, 3/4]$ there exists a constant $C_1 = C_1(x, p)$ so that*

$$(2.9) \quad \lim_{n \rightarrow \infty} \mathbb{P} \left\{ R_{p^{-1}n - xn^a, n}^{(\text{ind})} > C_1 n^{2a-1} \right\} = 0.$$

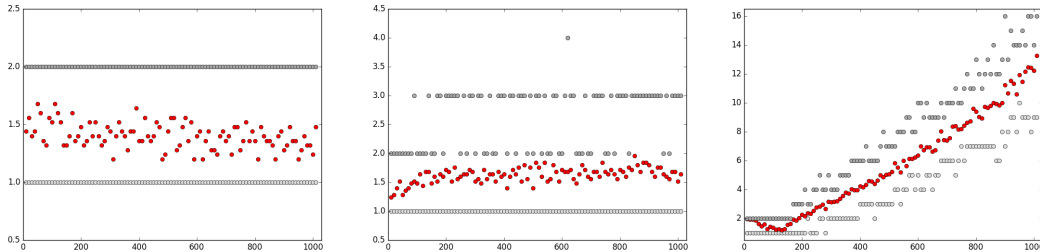


FIGURE 4. Monte Carlo simulations for the empirical maximum, minimum and expected number of regions for up to $n = 1000$ in the independent model for $a = 0.8$ with varying $p = 0.05, 0.5, 0.8$ from left to right. For each n , 25 independent environments were sampled. n grows in increments of size 10

(3) If $a \in (3/4, 1)$ there exists a constant $C_2 = C(x, p)$ so that,

$$(2.10) \quad \lim_{n \rightarrow \infty} \mathbb{P} \left\{ R_{p^{-1}n - xn^a, n}^{(\text{ind})} > C_2 n^{2a/3} \right\} = 0.$$

Finally when $a \in [1/2, 5/7]$ we obtain Tracy-Widom fluctuations. These fluctuations, instead of the standard way of centering the random variable by the mean and scaling by the variance, they manifest when we deterministically alter the dimensions of the rectangle by a smaller order term. In particular we prove

THEOREM 2.8. For $s \in \mathbb{R}$ define $x = \frac{2}{\sqrt{p}} \left(\frac{q}{p} \right)^a$ and $y(s) = s \frac{\sqrt{p}}{q} \left(\frac{p}{q} \right)^{\frac{1+a}{3}}$. Then

(1) For $1/2 < a < 2/3$,

$$(2.11) \quad \lim_{n \rightarrow \infty} \mathbb{P} \left\{ G_{\frac{n}{p} - xn^a - y(s)n^{\frac{2-a}{3}}, n}^{(0)} \leq n - \left(\frac{qn}{p} \right)^{2a-1} \right\} = F_{TW}(s).$$

(2) For $2/3 \leq a \leq 5/7$,

$$(2.12) \quad \lim_{n \rightarrow \infty} \mathbb{P} \left\{ G_{\frac{n}{p} - xn^a - y(s)n^{\frac{2-a}{3}}, n}^{(0)} \leq n - \left(\frac{qn}{p} \right)^{2a-1} + ax \left(\frac{q}{p} \right)^{2a-2} n^{3a-2} \right\} = F_{TW}(s).$$

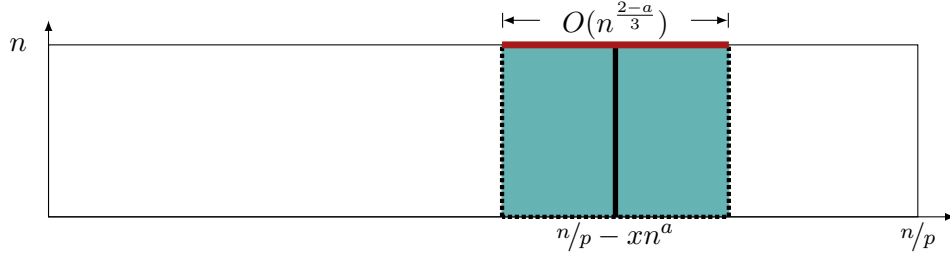


FIGURE 5. Tracy-Widom fluctuations to the last passage time of the independent model depend on position of the endpoint in the thickset red line. When $a \in (1/2, 2/3)$ the Tracy-Widom reveals itself just by centering according to the first and second order macroscopic approximation of the LLN for G . However when $a \in (3/2, 5/7)$, a third order approximation to the LLN, cn^{3a-2} is necessary in order to see the Tracy-Widom fluctuations.

Remark 2.9. The case $a \geq 5/7$ corresponds to an exponent $\gamma = 2a - 1 \geq 3/7$ in equation (1.5) (see [7]) and the result cannot be extended further with these techniques. In Section 3.1 of [7] the authors explain why their result should extend at least up to exponent $\gamma = 3/4$. The equivalent independent model here may be a bit more sensitive to these cut-offs and indeed $\gamma = 3/7$ manifests itself in the proof.

First from the last two cases of Theorem 2.8 we see that we need to amend the RHS in the probability by a term $O(n^{3a-2})$ and this gives a new cut-off $a = 2/3$ (or $\gamma = 1/3$). The term is there for case 2 as well, but when $a \leq 2/3$ the term is bounded and plays no role, while it must be dealt with, for higher a .

Second, from the proof of Theorem 2.8, the exponent $a = 5/7$ ($\gamma = 3/7$) seems to be critical, since it is necessary to have $2a - 1 < \frac{2-a}{3}$ to balance the various orders of magnitude that

appear. Assuming that the scaling in (1.5) remains the same for $\gamma \in (3/7, 3/4)$, this change implies a corresponding correction term of size $O(n^\gamma)$ at the numerator of (1.5).

3. OPTIMALITY REGIONS: MODEL-INDEPENDENT

We first present preliminary results about our models that do not depend on the correlation structure of the weights. We therefore write $R_{m,n}$ to mean either $R_{m,n}^{(\text{al})}$ or $R_{m,n}^{(\text{ind})}$.

3.1. Combinatorics. Let $\pi = \{u_0, \dots, u_m\} \in \Pi$ and recall that the increments $z_k = u_k - u_{k-1} \in \mathcal{R} = \{e_1, e_2, e_1 + e_2\}$. For each increment there are three possibilities:

- (1) $z_k = e_1 + e_2$ with $\omega_{u_k} = 0$, called a *mismatch*,
- (2) $z_k \in \{e_1, e_2\}$, called a *gap*,
- (3) $z_k = e_1 + e_2$ with $\omega_{u_k} = 1$, called a *match*.

Let $x = x(\pi)$ be the number of mismatches, $y = y(\pi)$ the number of gaps and $z = z(\pi)$ the number of matches of π . We also denote this triplet by

$$\mathbf{s}(\pi) = (x(\pi), y(\pi), z(\pi)).$$

Fix parameters $\alpha, \beta \geq 0$. The score of the path π is then given by

$$(3.1) \quad w_{\alpha,\beta}(\pi) = z - \alpha x - \beta y.$$

Since any diagonal step is equivalent to an e_1 step followed by a e_2 step or vice versa, we have

$$(3.2) \quad m + n = 2x(\pi) + 2z(\pi) + y(\pi) \quad \forall \pi \in \Pi.$$

The last passage time $G_{m,n}^{(\alpha,\beta)}$ (or $L_{m,n}^{(\alpha,\beta)}$, depending on the environment) defined in (1.2) can now be rewritten as

$$G_{m,n}^{(\alpha,\beta)} = \max_{\pi \in \Pi_{0,(m,n)}} \{w_{\alpha,\beta}(\pi)\}.$$

Our focus will be on the *minimal-gap maximisers* (MGM): paths whose score attains the last passage time with the smallest possible number of gaps. Since any two MGM paths have the same number of gaps and the same score it follows from (3.2) that

LEMMA 3.1. *All MGM paths have the same number of gaps, matches and mismatches.*

We denote the set of MGM paths by $\Gamma_{0,(m,n)}^{(\alpha,\beta)}$, or $\Gamma_{0,(m,n)}^{(\beta)}$. When $\alpha = 0$ we write $\Gamma_{0,(m,n)}^{(\beta)}$.

Definition 3.2. *Two points (α_1, β_1) and (α_2, β_2) belong in different optimality regions of the parameter space for a fixed terminal point (m, n) if and only if $\Gamma_{0,(m,n)}^{(\alpha_1, \beta_1)} \cap \Gamma_{0,(m,n)}^{(\alpha_2, \beta_2)} = \emptyset$.*

For future reference we record the following observations:

- (1) Fix $\alpha \geq 0$ then for any $\beta_1 \leq \beta_2$ we have

$$(3.3) \quad w_{\alpha,\beta_1}(\pi) \geq w_{\alpha,\beta_2}(\pi)$$

and therefore the inequality also holds for the passage times:

$$(3.4) \quad G_{m,n}^{(\alpha,\beta_1)} \geq G_{m,n}^{(\alpha,\beta_2)} \quad \text{and} \quad L_{m,n}^{(\alpha,\beta_1)} \geq L_{m,n}^{(\alpha,\beta_2)}$$

(2) For $\alpha = -1$ and $\beta = -1/2$, the weight of any path $\pi \in \Pi_{0,(m,n)}$ is given by

$$(3.5) \quad w_{-1,-1/2}(\pi) = \frac{m+n}{2}$$

LEMMA 3.3. *All optimality regions in the (α, β) -positive quadrant are semi-infinite cones bounded by the coordinate axes and lines of the form $\beta = c + \alpha(c + 1/2)$.*

This result was first proved in [19], we give a simplified proof here:

Proof. Pick any $(\alpha, \beta) \in \mathbb{R}_+^2$ and let $(0, \beta')$ be the point of intersection of the linear segment connecting (α, β) and $(-1, -1/2)$ with the y -axis, i.e.

$$(3.6) \quad \beta = (\alpha + 1)\beta' + \frac{\alpha}{2}.$$

We will show that the optimal paths associated with $(0, \beta')$ are the same as those associated to (α, β) . Consider any $\pi \in \Pi_{0,(m,n)}$ with $\mathbf{s}(\pi) = (x, y, z)$. Then

$$\begin{aligned} w_{\alpha,\beta}(\pi) &= z - x\alpha - y\beta = z - x\alpha - y\beta + y\beta' - y\beta' \\ &= w_{0,\beta'}(\pi) - x\alpha - (\beta - \beta')y \\ &= w_{0,\beta'}(\pi) - x\alpha - \left((\alpha + 1)\beta' + \frac{\alpha}{2} - \beta'\right)y, \quad \text{by (3.6),} \\ &= w_{0,\beta'}(\pi) - \alpha(m + n - w_{0,\beta'}(\pi)), \quad \text{by (3.2),} \\ (3.7) \quad &= (1 + \alpha)w_{0,\beta'}(\pi) - \alpha(m + n). \end{aligned}$$

So the weight of any path with parameters (α, β) is an affine function of the weight with parameters $(0, \beta')$ and the two parameters must belong to the same optimality region. \square

Under a fixed environment ω , we define the *critical penalties*

$$(3.8) \quad 0 < \beta_1 < \dots < \beta_{R_{m,n}} < \infty$$

to be the gap penalties for $\alpha = 0$ at which the optimality region changes. We will also write β_∞ for the last threshold $\beta_{R_{m,n}}$.

LEMMA 3.4 (Critical penalties). *For each $k \leq R_{m,n}$ let $\pi^{(\beta_k)} \in \Gamma_{0,(m,n)}^{(\beta_k)}$, with $\mathbf{s}(\pi^{(\beta_k)}) = (x_{\beta_k}, y_{\beta_k}, z_{\beta_k})$. Then*

$$(3.9) \quad \beta_{k+1} = \frac{z_{\beta_k} - z_{\beta_{k+1}}}{y_{\beta_k} - y_{\beta_{k+1}}}.$$

Proof. Continuity of the optimal score in the parameter β implies that at β_{k+1} the weights will be the same whether β_{k+1} is approached by above (considering scores of paths in $\Gamma_{0,(m,n)}^{(\beta_{k+1})}$) or from below (scores of paths in $\Gamma_{0,(m,n)}^{(\beta_k)}$). Therefore

$$z_{\beta_k} - \beta_{k+1}y_{\beta_k} = z_{\beta_{k+1}} - \beta_{k+1}y_{\beta_{k+1}}$$

which yields the conclusion. \square

Upper bounds in probability for the maximal value of $R_{m,n}$ can be found in [11]. For the LCS model these are sharp when the alphabet size grows to infinity. The results and

arguments in [11] can be extended to give the following upper bound that holds in any fixed environment and n large enough.

$$(3.10) \quad R_{\lfloor nx \rfloor + o(n), \lfloor ny \rfloor + o(n)} \leq Cn^{2/3}.$$

LEMMA 3.5. *For $\beta_0 = 0$ and each critical β_k in (3.8), choose an MGM path $\pi_k \in \Gamma_{0, (m, n)}^{(\beta_k)}$ with $\mathbf{s}(\pi_k) = (x_k, y_k, z_k)$ for $0 \leq k \leq R_{m, n}$. Then*

$$(3.11) \quad R_{m, n} \leq \min \left\{ z_0 - z_{R_{m, n}}, \frac{x_{R_{m, n}} - x_0}{2}, \frac{y_0 - y_{R_{m, n}}}{2}, n \wedge m - z_0 \right\}.$$

Proof. Distinct paths π_i differ in the number of diagonal steps and the number of gaps. Since a diagonal step is equivalent to two gaps, we have

$$y_i - y_{i+1} \geq 2.$$

Furthermore, it must be the case that

$$z_i - z_{i+1} \geq 1;$$

otherwise π_i would violate the MGM condition. By the last two equations and (3.2) we have

$$x_{i+1} - x_i \geq 2.$$

Adding over i gives the first three terms in the minimum of (3.11). For the last term note that $y_{R_{m, n}} = n \vee m - m \wedge n$. Since $x_0 \geq 0$ (3.2) yields $2(n \wedge m - z_0) \geq y_0 - y_{R_{m, n}}$. \square

Remark 3.6. Notice that the last bound in (3.11) can be written as

$$(3.12) \quad R_{m, n}^{(\text{al})} \leq n - L_{m, n}^{(0)} \quad \text{and} \quad R_{m, n}^{(\text{ind})} \leq n - G_{m, n}^{(0)}.$$

4. OPTIMALITY REGIONS IN THE ALIGNMENT MODEL

In this section we prove our results about the alignment model. Because of Lemma 3.3 and (3.7) it is enough to consider the case where $\alpha = 0$.

We construct a path with a score that is near-optimal under any penalty β and which attempts to minimise as much as possible the number of vertical steps. This will be important when we need a lower bound for the passage time under penalty β_R , where we know that the optimal path takes no vertical steps. Consider the following strategy (S) to create a path π_S :

- (1) If $a \leq 1/2$: For some appropriate constant c_1 (to be determined), move with $e_1 + e_2$ steps from 0 up to a fixed point $u_n(a) = (\lfloor \sqrt{c_1 n \log n} \rfloor, \lfloor \sqrt{c_1 n \log n} \rfloor)$.
- (2) If $a > 1/2$: For some constant c_2 (to be determined), move with $e_1 + e_2$ steps from 0 up to a fixed point $u_n(a) = (\lfloor |\mathcal{A}|^{-1} x n^a - \sqrt{c_2 n \log n} \rfloor, \lfloor |\mathcal{A}|^{-1} x n^a - \sqrt{c_2 n \log n} \rfloor)$.
- (3) Now, from $u_n(a)$ construct the path as follows
 - (a) If the path is on site (i, j) with $j < n$ and $\omega_{i+1, j+1} = 1$ then move diagonally with an $e_1 + e_2$ step, and now the path is on site $(i + 1, j + 1)$.
 - (b) If the path is on site (i, j) with $i < \lfloor |\mathcal{A}| n - x n^a \rfloor$ and $\omega_{i+1, j+1} = 0$ then move horizontally with an e_1 step, and now the path is on site $(i + 1, j)$.
 - (c) If $j = n$ or $i = \lfloor |\mathcal{A}| n - x n^a \rfloor$, move to $(\lfloor |\mathcal{A}| n - x n^a \rfloor, n)$.

Constants c_1, c_2 will be chosen so that the path constructed this way will eventually exit the north boundary and therefore take no vertical steps. Random variables

(4.1)

$$Y_j = |\{i \in \mathbb{N} : (i, j + u_n(a) \cdot e_2) \in \pi_S\}| \sim \text{Geom}\left(\frac{1}{|\mathcal{A}|}\right), \quad \mathbb{P}\{Y_j = \ell\} = 1/|\mathcal{A}|(1 - 1/|\mathcal{A}|)^{\ell-1}$$

give the length of the path π_S on any horizontal line after point u_n . By construction, they are independent.

We will frequently use the following moderate deviations result [9]:

LEMMA 4.1. *Let $(X_N)_{N \in \mathbb{N}}$ an i.i.d. sequence of random variables with exponential moments. If $N\lambda_N^2 \rightarrow \infty$ and $N\lambda_N^3 \rightarrow 0$ then*

$$\mathbb{P}\left\{\left|\frac{1}{N} \sum_{i=1}^N X_i - \mathbb{E}(X_1)\right| > \lambda_N\right\} \sim \frac{2}{\sqrt{2\pi N\lambda_N^2}} e^{-N\lambda_N^2/2}.$$

4.1. Proof of Theorem 2.1. If the path π_S hits the north boundary first, it may conceivably be an optimal path under β_R and utilise the last passage time. We will choose the constants c_1, c_2 so that this does not happen with summable probabilities. We consider the case $a \leq 1/2$. Since the proof for $a > 1/2$ is similar it is left for the reader. The probability that we hit the north boundary is given by

$$\begin{aligned} & \mathbb{P}\left\{\sum_{i=1}^{n-u_n(a) \cdot e_2} Y_i \leq \lfloor n|\mathcal{A}| - xn^a \rfloor - \lfloor \sqrt{c_1 n \log n} \rfloor\right\} \\ & \geq \mathbb{P}\left\{\sum_{i=1}^{n-\lfloor \sqrt{c_1 n \log n} \rfloor} (Y_i - |\mathcal{A}|) \leq -xn^a + (|\mathcal{A}| - 1)\sqrt{c_1 n \log n} - 3\right\} \\ & \geq 1 - c_0 \frac{1}{\sqrt{\log n}} n^{-c_1(|\mathcal{A}|-1)^2/2}, \end{aligned}$$

where the last inequality follows from Lemma 4.1. For $c_1 > 2/(|\mathcal{A}|-1)^2$, the complementary probability is summable, so by the Borel-Cantelli lemma we conclude that almost surely there exists random $N \in \mathbb{N}$ such that for $n > N(\omega)$ the path π_S hits from the north boundary. When that happens we can estimate the total number of gaps:

$$y(\pi_S) = \sum_{i=1}^{n-\lfloor \sqrt{c_1 n \log n} \rfloor} Y_i - n + \lfloor \sqrt{c_1 n \log n} \rfloor \leq n(|\mathcal{A}| - 1) + o(\sqrt{n \log n}), \quad n > N(\omega),$$

by the law of large numbers and Lemma 4.1.

Next estimate the number of matches on π_S . For the initial segment of the path, when the increments are all $e_1 + e_2$ the path collects $1/|\mathcal{A}|(\sqrt{c_1 n \log n}) + o(\sqrt{n \log n})$ many points. After site $u_n(a)$, it collects a point from every horizontal line until it exits. Overall, and using again Lemma 4.1, there exists $M(\omega) \geq N(\omega)$ such that,

$$z(\pi_S) = n - \left(1 - \frac{1}{|\mathcal{A}|}\right) \sqrt{c_1 n \log n} + o(\sqrt{n \log n}), \quad n > M(\omega).$$

This path may be optimal under any penalty, therefore for $n > M(\omega)$, the score is bounded by

$$L_{[n|\mathcal{A}|-xn^a],n}^{(\beta)} \geq n(1 + \beta - |\mathcal{A}|\beta) - \left(1 - \frac{1}{|\mathcal{A}|}\right) \sqrt{c_1 n \log n} + o(\sqrt{n \log n}), \quad 0 \leq a \leq 1/2.$$

This completes the proof.

4.2. Proof of Theorem 2.2. Let

$$g(n) = \begin{cases} \sqrt{n \log n}, & a \leq 1/2, \\ n^a, & a > 1/2. \end{cases}$$

The bound of Theorem 2.1 also gives a bound on the number of gaps when $\beta = 0$: Matches are obtained via diagonal steps, therefore the optimal path takes more than $n - C_1 g(n)$ diagonal steps. By (3.2), the number of non-diagonal steps of the optimal path is less than $n/p - n - xn^a + C_2 g(n)$. At penalty β_R the path takes precisely $n/p - n - xn^a$ non-diagonal steps and therefore $y_0 - y_R < C_3 g(n)$. Furthermore, since Theorem 2.1 holds for any penalty β , we also have that eventually

$$(4.2) \quad z_0 - z_R < C_4 g(n).$$

Therefore, setting $C = C_3 \vee C_4$,

$$z_0 - z_R + y_0 - y_R = \sum_{i=0}^{R-1} \{(z_i - z_{i+1}) + (y_i - y_{i+1})\} \leq C g(n).$$

Now we are in a position to use the techniques in [11]. The main idea is as follows: The sum above has as terms the numerators and denominators of the critical penalties (see Lemma 3.4). Each critical penalty is a distinct rational number and it corresponds to a change of optimality region. The LHS bound is environment-independent, so we can obtain an upper bound on the number of regions if maximise the number of terms in the sum. Take each successive integer k and compute the number of irreducible fractions a/b so that $a + b = k$. The number of irreducible fractions satisfying this is $\varphi(k)$, where φ is Euler's totient function [3] and therefore, in our case, the maximal possible number of summands is the integer R_{\max} satisfying

$$\sum_{i=1}^{R_{\max}} k \varphi(k) \leq g(n) < \sum_{i=1}^{R_{\max}+1} k \varphi(k).$$

These inequalities imply that R_{\max} will be bounded above, up to a lower order term, by $C g(n)^{2/3}$. This follows by the asymptotics of φ for large arguments, and we direct the reader to the proof of Theorem 5 in [11] for the details. The result for the optimality regions follows since $R_{m_n,n}^{(\text{al})} \leq R_{\max}$.

4.3. Proof of Theorem 2.3. Again, we present the case $a \leq 1/2$. Apply strategy (S) with $u_n(a) = (0, 0)$ so that part (3) applies immediately. Recalling the Y_j from (4.1)

$$\mathbb{E} \left[L_{[n|\mathcal{A}|-xn^a],n}^{(0)} \right] = \sum_{k=0}^n \mathbb{P} \left\{ L_{[n|\mathcal{A}|-xn^a],n}^{(0)} \geq k \right\} \geq \sum_{k=0}^n \mathbb{P} \left\{ \sum_{j=0}^k Y_j \leq [n|\mathcal{A}| - xn^a] \right\}$$

$$\begin{aligned}
&\geq \sum_{k=0}^{n-\sqrt{2/|\mathcal{A}|n \log n}} \mathbb{P} \left\{ \sum_{j=0}^k Y_j \leq \lfloor n|\mathcal{A}| - xn^a \rfloor \right\} \\
&\quad + \sum_{k=n-\sqrt{2/|\mathcal{A}|n \log n}+1}^{n-\sqrt{1/|\mathcal{A}|n \log(\log n)}} \mathbb{P} \left\{ \sum_{j=0}^k Y_j \leq \lfloor n|\mathcal{A}| - xn^a \rfloor \right\} \\
&\geq (n - \sqrt{2/|\mathcal{A}|n \log n}) (1 - C_0 \frac{1}{\sqrt{\log n}} n^{-1}) \\
&\quad + (\sqrt{2/|\mathcal{A}|n \log n} - \sqrt{2/|\mathcal{A}|n \log(\log n)}) (1 - C_1 \frac{1}{\sqrt{\log(\log n)}} (\log n)^{-1/2}) \\
&= n - \sqrt{2/|\mathcal{A}|n \log(\log n)} - o(\sqrt{n}).
\end{aligned}$$

where the third inequality once more follows from Lemma 4.1. It now suffices to apply the bound $R_{m,n}^{(\text{ind})} < n - z_0$ from equation (3.11) to conclude the result.

5. OPTIMALITY REGIONS FOR THE INDEPENDENT MODEL

In this section we prove results about optimality regions in the independent model. The following identity couples the independent model with the corner growth model in an i.i.d. $\text{Geom}(1-p)$ environment. Recall that $T_{m,n}$ denotes the last passage time in an $m \times n$ rectangle, with admissible e_1 or e_2 steps only.

$$(5.1) \quad \mathbb{P} \left\{ G_{m,n}^{(0)} \leq m - N \right\} = \mathbb{P} \{ T_{n-m+N,N} \leq n + N - 1 \}.$$

The result follows from the arguments in [13], and we briefly present the main idea.

The *discrete totally asymmetric simple exclusion process* (DTASEP) with backward updating is an interacting particle system of left-finite particle configuration on the integer lattice, i.e. such that sites to the left of some threshold are empty. Label the particles from left to right and denote the position of the j^{th} particle at time $\ell \in \mathbb{N}$ by $\eta_j(\ell)$. At every discrete time step $\ell \in \mathbb{N}$ each particle independently attempts to jump one step to the left with probability $q = 1 - p$. Particle i performs the jump if either

- (1) the target site was unoccupied by particle $i - 1$ at time $\ell - 1$ or,
- (2) the target site was occupied by particle $i - 1$, but it also performs a jump at time ℓ .

In words, particles are forbidden to jump to occupied sites and we update from left to right. Start DTASEP with the *step initial condition* $\eta_i(0) = i$ so that initially the i -th particle is at position i . Let $\tau_{i,j}$ be the time it takes particle j to jump i times:

$$\tau_{i,j} = \inf \{ \ell \geq 0 : \eta_j(\ell) \leq j - i \}.$$

Then the following recursive equation holds

$$\tau_{i,j} = \tau_{i,j-1} \vee (\tau_{i-1,j} + 1) + \tilde{\zeta}_{i,j}.$$

where the $\tilde{\zeta}_{i,j}$ are independent Geometric variables with parameter p .

By setting $\zeta_{i,j} = \tilde{\zeta}_{i,j} + 1$ the $\tau_{i,j}$ can be coupled with the last passage time in the corner growth model (cf. [13], Lemma 5.1), giving the equality in distribution

$$(5.2) \quad \tau_{i,j} \stackrel{(d)}{=} T_{i,j} - j + 1.$$

We embed DTASEP in the two-dimensional lattice $\mathbb{Z} \times \mathbb{N}_+$, using its graphical construction. Assign on each site (k, ℓ) of the lattice an i.i.d. Bernoulli(q) weight $\zeta_{k,\ell}$. Particles are placed initially on $\mathbb{N}_+ \times \{0\}$, with particle i at coordinate $(\eta_i(0), 0)$. The Bernoulli marked sites signify which particles will attempt to jump in the DTASEP process. After the spatial locations in the DTASEP at time $\ell = 1$ are determined, the particles in the graphical construction are at positions $(\eta_i(1), 1)$. We iterate this procedure for all times $\ell \in \mathbb{N}$. Then

$$1 - \omega_{k,\ell} = \zeta_{k+\ell,\ell}.$$

In [13] the following combinatorial identity was proved:

$$(5.3) \quad G_{m,n}^{(0)} = m - \max\{k : (m-n) \vee 1 \leq k \leq m, \tau_{k+n-m,k} \leq n\}.$$

Set $k^* = \max\{k \leq m : k \geq (m-n) \vee 1, \tau_{k+n-m,k} \leq n\} \vee 0$. Then

$$(5.4) \quad \{G_{m,n}^{(0)} \leq m - N\} = \{N \leq k^*\} = \{\tau_{N+n-m,N} \leq n\},$$

where the last equality comes from the fact that $\tau_{N+n-m,N}$ is an increasing random variable in N . Finally compute

$$\begin{aligned} \mathbb{P}\{G_{m,n}^{(0)} \leq m - N\} &= \mathbb{P}\{N \leq \max\{k : (m-n) \vee 1 \leq k \leq m, \tau_{k+n-m,k} \leq n\}\}, \quad \text{by (5.3)} \\ &= \mathbb{P}\{\tau_{N+n-m,N} \leq n\}, \quad \text{by (5.4)} \\ &= \mathbb{P}\{T_{N+n-m,N} \leq n + N - 1\}, \quad \text{by (5.2)}. \end{aligned}$$

5.1. Proof of Theorem 2.5. Recall that $m_n = n/p - xn^a$ and $a \in (0, 1/2)$. Our goal is to prove that the sequence of random variables $n - G_{n,m_n}^{(0)}$ is tight. The main ingredient in the proof is the identity (5.1). Set $N = \frac{nq}{p} - xn^a + k$. Then

$$(5.5) \quad n - m + N = n - \frac{n}{p} + xn^a + \frac{nq}{p} - xn^a + k = k.$$

Since $N(n)$ is eventually monotone, we can invert the expression above and find n in terms of N for sufficiently large n (and hence N). In particular,

$$(5.6) \quad n = n(N) = \frac{p}{q}N + xN^a \left(\frac{p}{q}\right)^{a+1} + O(N^{2a-1}).$$

To see this we compute

$$\begin{aligned} N(n(N)) &= \frac{q}{p}n(N) - xn(N)^a + k \\ &= \frac{q}{p} \left(\frac{p}{q}N + xN^a \left(\frac{p}{q}\right)^{a+1} + O(N^{2a-1}) \right) \end{aligned}$$

$$\begin{aligned}
& -x \left(\frac{p}{q} N + x N^a \left(\frac{p}{q} \right)^{a+1} + O(N^{2a-1}) \right)^a + k \\
& = N + \left(\frac{p}{q} \right)^a x N^a - x \left(\frac{p}{q} N \right)^a \left(1 + x N^{a-1} \left(\frac{p}{q} \right)^a + O(N^{2a-2}) \right)^a + O(1) \\
& = N + \left(\frac{p}{q} \right)^a x N^a - x \left(\frac{p}{q} N \right)^a \left(1 + a x N^{a-1} \left(\frac{p}{q} \right)^a + O(N^{2a-2}) \right) + O(1) \\
& = N + O(1).
\end{aligned}$$

Therefore, $n + N - 1 = \frac{N}{q} + x \left(\frac{p}{q} \right)^{a+1} N^a + O(N^{2a-1})$. Combine in (5.1)

$$\begin{aligned}
\mathbb{P}\{k \leq n - G_{m_n, n}^{(0)}\} &= \mathbb{P}\{G_{m_n, n}^{(0)} \leq m_n - N\} \\
&= \mathbb{P}\left\{T_{k, N} \leq \frac{N}{q} + x \left(\frac{p}{q} \right)^{a+1} N^a + O(N^{2a-1})\right\} \\
&\leq \mathbb{P}\left\{\max_{j: 1 \leq j \leq k} \sum_{i=1}^N \zeta_{i, j} \leq \frac{N}{q} + x \left(\frac{p}{q} \right)^{a+1} N^a + O(N^{2a-1})\right\} \\
(5.7) \quad &= \mathbb{P}\left\{\sum_{i=1}^N \zeta_{i, 1} - N\mathbb{E}(\zeta_{11}) \leq x \left(\frac{p}{q} \right)^{a+1} N^a + O(N^{2a-1})\right\}^k.
\end{aligned}$$

The results follow by first dividing by $\frac{\sqrt{p}}{q}\sqrt{N}$ and the central limit theorem, when we let n (hence N) tend to infinity.

5.2. Proof of Theorem 2.6. We first show the result when $a < 1/2$. Using equations (3.12) from Remark 3.6 and (5.7) from the proof of Theorem 2.5, we have

$$\begin{aligned}
\mathbb{P}\{k < R_{\frac{n}{p} - xn^a, n}^{(\text{ind})}\} &\leq \mathbb{P}\{k < n - G_{\frac{n}{p} - xn^a, n}^{(0)}\} \\
(5.8) \quad &= \left(\mathbb{P}\left\{\frac{\sum_{i=1}^N \zeta_{i, 1} - \mathbb{E}(\zeta_{i, 1})N}{\sqrt{\text{Var}(\zeta_{i, 1})N}} < C_1 N^{a-1/2}\right\} \right)^k
\end{aligned}$$

for C_1 large enough. As in the proof of Theorem 2.5 we have $N = \frac{nq}{p} - xn^a + k$ and let Φ denote the cumulative distribution function of the standard normal distribution. Fix a tolerance $\delta > 0$ satisfying $\Phi(\delta) + \delta < 1$ and let $n_1(\delta)$ large enough so that $C_1 N^{a-1/2} < \delta$ for all $n > n_1(\delta)$. Applying the Berry-Esseen theorem to the last line of the last display,

$$(5.9) \quad \mathbb{P}\{k \leq R_{\frac{n}{p} - xn^a, n}^{(\text{ind})}\} \leq \left(\Phi(\delta) + \frac{C}{\sqrt{n}} \right)^k \leq (\Phi(\delta) + \delta)^k, \quad \text{for all } n > n_2(\delta).$$

For $n \geq n_0(\delta) = n_1(\delta) \vee n_2(\delta)$ the right hand side of (5.9) is uniformly summable in k . Moreover, by (5.9) and the reverse Fatou's Lemma we compute

$$\overline{\lim}_{n \rightarrow \infty} \mathbb{E} \left[R_{\frac{n}{p} - xn^a, n}^{(\text{ind})} \right] = \overline{\lim}_{n \rightarrow \infty} \sum_{k=0}^{\infty} \mathbb{P}\{k \leq R_{\frac{n}{p} - xn^a, n}^{(\text{ind})}\}$$

$$\leq \sum_{k=0}^{\infty} \overline{\lim}_{n \rightarrow \infty} \mathbb{P}\{k < n - G_{\frac{n}{p} - xn^a, n}^{(0)}\} \leq \sum_{k=0}^{\infty} 2^{-k} = 2,$$

where the penultimate inequality follows from (3.12) and the last from Theorem 2.5.

The case $a = 1/2$ is slightly more delicate, but the ideas are exactly the same. As before,

$$(5.10) \quad \mathbb{P}\{k < R_{\frac{n}{p} - x\sqrt{n}, n}^{(\text{ind})}\} \leq \left(\mathbb{P}\left\{ \frac{\sum_{i=1}^N \zeta_i - \mathbb{E}(\zeta_1)N}{\sqrt{\text{Var}(\zeta_1)N}} < x \frac{p^{5/2}}{q^2} + C_0 N^{-1/2} \right\} \right)^k.$$

The right-hand side converges to $(\Phi(xp^{5/2}q^{-2}))^k$ and with the same arguments as before,

$$\overline{\lim}_{n \rightarrow \infty} \mathbb{E} \left[R_{\frac{n}{p} - xn^a, n}^{(\text{ind})} \right] \leq \frac{1}{1 - \Phi(xp^{5/2}q^{-2})}.$$

5.3. Proof of Theorem 2.7. When $a \leq 1/2$ the result is immediate from equations (5.8), (5.10). For $a \in [1/2, 3/4]$

$$\begin{aligned} \overline{\lim}_{n \rightarrow \infty} \mathbb{P}\left\{ \left(\frac{(px)^2}{4(1-p)} + \varepsilon \right) n^{2a-1} \leq R_{p^{-1}n - xn^a, n}^{(\text{ind})} \right\} \\ \leq \overline{\lim}_{n \rightarrow \infty} \mathbb{P}\left\{ \left(\frac{(px)^2}{4(1-p)} + \varepsilon \right) n^{2a-1} \leq n - G_{p^{-1}n - xn^a, n}^{(0)} \right\} = 0. \end{aligned}$$

The last equality follows from (1.6). This gives the second part of the statement.

When $a \in (3/4, 1)$ we can obtain a sharper bound using the technique in [11]. From the proof of Lemma 3.5 and then (4.2) we can find a constant C_1 (that may depend on x) such that $n - G_{p^{-1}n - xn^a, n}^{(\beta_R)} = n - z_R < C_1 n^a$ for n large enough. Therefore, for n large enough

$$(5.11) \quad z_0 - z_R < C_1 n^a.$$

Moreover, since the number of vertical steps at $\beta = 0$ cannot exceed $n - G_{p^{-1}n - xn^a, n}^{(0)}$, (1.6) gives that with probability tending to 1

$$(5.12) \quad y_0 - y_R \leq n - G_{p^{-1}n - xn^a, n}^{(0)} < C_2 n^{2a-1}.$$

Equations (5.11), (5.12) now yield a constant C such that

$$(5.13) \quad \lim_{n \rightarrow \infty} \mathbb{P}\{z_0 - z_R + y_0 - y_R < Cn^a\} = 1.$$

Let A_n the event in the probability above. On A_n ,

$$\sum_{i=0}^{R-1} \{(z_i - z_{i+1}) + (y_i - y_{i+1})\} < Cn^a.$$

Now we once more follow the approach of [11] as in the proof of Theorem 2.2.

6. FLUCTUATIONS FOR BLIP

Here we prove Theorem 2.8. We will once more use (5.1). Recall that

$$x = \frac{2}{\sqrt{p}} \left(\frac{q}{p}\right)^a \quad \text{and} \quad y = s \frac{\sqrt{p}}{q} \left(\frac{p}{q}\right)^{\frac{1+a}{3}}, \quad s \in \mathbb{R}.$$

We further define

$$(6.1) \quad N = N(n) = \frac{q}{p}n - xn^a - yn^{\frac{2-a}{3}} + c_n,$$

where c_n is given by

$$(6.2) \quad c_n = \begin{cases} \left(\frac{q}{p}\right)^{2a-1} n^{2a-1}, & 1/2 < a < 2/3, \\ \left(\frac{q}{p}\right)^{2a-1} n^{2a-1} - (2a-1)x \left(\frac{q}{p}\right)^{2a-2} n^{3a-2}, & 2/3 \leq a < 5/7. \end{cases}$$

Observe that $N(n)$ in equation (6.1) is an eventually monotone function. Therefore, for N large enough, there is a well defined inverse $n = n(N)$. We define

$$\ell(N) = \frac{p}{q}N + \frac{2\sqrt{p}}{q}N^a + y \left(\frac{q}{p}\right)^{\frac{1+a}{3}} N^{\frac{2-a}{3}}.$$

A Taylor expansion yields

$$(6.3) \quad |N - N(\ell(N))| = |N - \frac{q}{p}\ell(N) + x_{p,a}\ell(N)^a + y\ell(N)^{\frac{2-a}{3}}| = O(N^{2a-1}).$$

and therefore $|n(N) - \ell(N)| = o(N^{\frac{2-a}{3}})$; in fact we will show that

$$(6.4) \quad |n(N) - \ell(N)| < cN^\beta.$$

Assume for a contradiction that (6.4) does not hold for some $c > 0$. Then

$$\begin{aligned} |N - N(\ell(N))| &= |N(n(N)) - N(\ell(N))| \\ &= \left| \frac{q}{p}(n(N) - \ell(N)) - x(n(N)^a - \ell(N)^a) \right. \\ &\quad \left. - y(n(N)^{\frac{2-a}{3}} - \ell(N)^{\frac{2-a}{3}}) + c_{n(N)} - c_{\ell(N)} \right| \\ &\geq \frac{q}{p}|n(N) - \ell(N)| - x|n(N) - \ell(N)|^a - |y||n(N) - \ell(N)|^{\frac{2-a}{3}} \\ &\quad - |c_{n(N)} - c_{\ell(N)}| \\ &\geq CN^\beta \text{ for some } C > 0 \text{ and } N \text{ large enough.} \end{aligned}$$

This contradicts (6.3) since $\beta > 2a - 1$. In particular we have shown that

$$(6.5) \quad \lim_{N \rightarrow \infty} \frac{n(N) - \frac{p}{q}N - \frac{2\sqrt{p}}{q}N^a - y \left(\frac{q}{p}\right)^{\frac{1+a}{3}} N^{\frac{2-a}{3}}}{N^{\frac{2-a}{3}}} = 0.$$

Next we prove $c_n = N^{2a-1} + o(1)$: for $a < 2/3$ this is immediate. When $a \in [2/3, 5/7)$,

$$(6.6) \quad N^{2a-1} = \left(\frac{q}{p}n\right)^{2a-1} \left(1 - (2a-1)\frac{px}{q}n^{a-1} + O\left(n^{-\frac{1+a}{3}}\right)\right) = c_n + O(n^{\frac{5a-4}{3}}).$$

For $a \in [2/3, 5/7)$ the exponent $\frac{5a-4}{3} < 0$, so $c_n = N^{2a-1} + o(1)$ follows. Therefore,

$$(6.7) \quad m - n = \frac{q}{p}n - xn^a - yn^{\frac{2-a}{3}} = N - c_n = N - N^{2a-1} + o(1).$$

To finish the proof we need to be a bit cautious with the integers parts. Define k_N to be

$$k_N = \lfloor m \rfloor - n - \lfloor N \rfloor + \lfloor \lfloor N \rfloor^{2a-1} \rfloor.$$

It follows from (6.7) that k_N is bounded in N (and n). Also set $N = \lfloor N \rfloor + \varepsilon_N$. Substituting these in equation (5.1) we compute

$$(6.8) \quad \begin{aligned} \mathbb{P}\{G_{\lfloor m \rfloor, n}^{(0)} \leq n - \lfloor N^{2a-1} \rfloor\} &= \mathbb{P}\{T_{\lfloor \lfloor N \rfloor^{2a-1} \rfloor, \lfloor N \rfloor + k_N} \leq n + \lfloor N \rfloor - 1\} \\ &= \mathbb{P}\{T_{\lfloor \lfloor N \rfloor^{2a-1} \rfloor, \lfloor N \rfloor + k_N} \leq \ell(N) + N - 1 + n - \ell(N) + \varepsilon_N\} \\ &= \mathbb{P}\left\{T_{\lfloor \lfloor N \rfloor^{2a-1} \rfloor, \lfloor N \rfloor + k_N} - \frac{1}{q}N - \frac{2\sqrt{p}}{q}N^a \leq y \left(\frac{q}{p}\right)^{\frac{1+a}{3}} N^{\frac{2-a}{3}} - 1 + n - \ell(N) + \varepsilon_N\right\} \\ &= \mathbb{P}\left\{\frac{T_{\lfloor \lfloor N \rfloor^{2a-1} \rfloor, \lfloor N \rfloor + k_N} - \frac{1}{q}N - \frac{2\sqrt{p}}{q}N^a}{\frac{\sqrt{p}}{q}N^{\frac{2-a}{3}}} \leq s + o(1)\right\}. \end{aligned}$$

The passage time in the probability above can be compared with $T_{\lfloor N^{2a-1} \rfloor, \lfloor N \rfloor}$ and satisfies

$$|T_{\lfloor \lfloor N \rfloor^{2a-1} \rfloor, \lfloor N \rfloor} - T_{\lfloor \lfloor N \rfloor^{2a-1} \rfloor, \lfloor N \rfloor + k_N}| < \sum_{i=0}^{\lfloor \lfloor N \rfloor^{2a-1} \rfloor} \sum_{j=-k_N}^{k_N} \zeta_{i, \lfloor N \rfloor + j}.$$

Since $a < 5/7$, the number of geometric random variables in the right-hand side of the inequality is of lower order than $N^{\frac{2-a}{3}}$ and when scaled by it the double sum vanishes \mathbb{P} -a.s. This allows us to remove k_N from (6.8) and equation (1.5) now yields the result by taking $n \rightarrow \infty$.

REFERENCES

- [1] ALURU, S., Ed. *Handbook of computational molecular biology*. Chapman & Hall/CRC Computer and Information Science Series. Chapman & Hall/CRC, Boca Raton, FL, 2006.
- [2] AMSALU, S., MATZINGER, H., AND VACHKOVSKAIA, M. Thermodynamical approach to the longest common subsequence problem. *J. Stat. Phys.* 131, 6 (2008), 1103–1120.
- [3] APOSTOL, T. M. *Introduction to Analytic Number Theory*, 5th edition ed. Undergraduate Texts in Mathematics. Springer, 1995.
- [4] BARYSHNIKOV, Y. GUEs and queues. *Prob. Theory Relat. Fields* 119 (2001), 256 – 274.
- [5] BASDEVANT, A.-L., ENRIQUEZ, N., GERIN, L., AND GOUÉRÉ, J.-B. Discrete Hammersley’s lines with sources and sinks. *ALEA Lat. Am. J. Probab. Math. Stat.* 13 (2016), 33–52.
- [6] BERGROTH, L., HAKONEN, H., AND RAITA, T. A survey of longest common subsequence algorithms. *SPIRE 00* (2000), 39 – 48.
- [7] BODINEAU, T., AND MARTIN, J. A universality property for last-passage percolation close to the axis. *Electron. Commun. Probab.* 10, 11 (2005), 105 – 112.
- [8] CHVÁTAL, V., AND SANKOFF, D. Longest common subsequences of two random sequences. *J. Appl. Probab.* 12, 2 (1975), 306 – 315.
- [9] CRAMÈR, H. Sur un nouveau théorème limite de la probabilité. *Actualités Sci. Industr.* 736 (1938), 5 – 23.

- [10] DEWEY, C. N., HUGGINS, P. M., WOODS, K., STURMFELS, B., AND PACTER, L. Parametric alignment of drosophila genomes. *PLoS Comput Biol* 2, 6 (2006), e73.
- [11] FERNÁNDEZ-BACA, D., SEPPÄLÄINEN, T., AND SLUTZKI, G. Bounds for parametric sequence comparison. *Discrete Appl. Math.* 118 (2002), 181 – 198.
- [12] FERNÁNDEZ-BACA, D., AND VENKATACHALAM, B. *Parametric sequence alignment*. CRC Press Computer and Information Science Series. Chapman and Hall, 2006.
- [13] GEORGIOU, N. Soft edge results for longest increasing paths on the planar lattice. *Electron. J. Probab.* 15 (2010), 1–13.
- [14] GEORGIOU, N., RASSOUL-AGHA, F., AND SEPPÄLÄINEN, T. Variational formulas and cocycle solutions for directed polymer and percolation models. *Commun. Math. Phys.* 346, 2 (September 2016), 741–779.
- [15] N. Georgiou, F. Rassoul-Agha, T. Seppäläinen. Stationary cocycles and Busemann functions for the corner growth model. *Probab. Theory Relat. Fields*, 2016. DOI: 10.1007/s00440-016-0729-x.
- [16] N. Georgiou, F. Rassoul-Agha, T. Seppäläinen. Geodesics and the competition interface for the corner growth model. *Probab. Theory Relat. Fields*, 2016 DOI:10.1007/s00440-016-0734-0.
- [17] GLYNN, P. W., AND WHITT, W. Departures from many queues in series. *Ann. Appl. Probab.* 1, 4 (1991), 546 – 572.
- [18] GONG, R., HOUDRÉ, C., AND LEMBER, J. Lower bounds on the generalized central moments of the optimal alignments score of random sequences. *ArXiv:1506.06067* (2015).
- [19] GUSFIELD, D., BALASUBRAMANIAN, K., AND NAOR, D. Parametric optimization of sequence alignment. *Algorithmica* 12, 4-5 (1994), 312–326.
- [20] HAMMERSLEY, J. M. A few seedlings of research. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. I: Theory of statistics* (1972), Univ. California Press, Berkeley, Calif., pp. 345–394.
- [21] HENIKOFF, S., AND HENIKOFF, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89, 22 (Nov. 1992), 10915 – 10919.
- [22] HIRSCHBERG, D. S. A linear space algorithm for computing maximal common subsequences. *Commun. ACM* 18, 6 (1975), 341 – 343.
- [23] HOUDRÉ, C., AND MATZINGER, H. Closeness to the diagonal for longest common subsequences in random words. *Electron. Commun. Probab.* 21, 36 (2016), 1 – 19.
- [24] HOWER, V., AND HEITSCH, C. E. Parametric analysis of RNA branching configurations. *Bull. Math. Biol.* 73, 4 (2011), 754 – 776.
- [25] KIWI, M., LOEBL, M., AND MATOUŠEK, J. Expected length of the longest common subsequence for large alphabets. *Adv. Math.* 197, 2 (2005), 480–498.
- [26] KOMLÓS, J., MAJOR, P., AND TUSNÁDY, G. An approximation of partial sums of independent RV's, and the sample DF. II. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 34 (1976), 33–58.
- [27] LEMBER, J., AND MATZINGER, H. Standard deviation of the longest common subsequence. *Ann. Probab.* 37, 3 (2009), 1192 – 1235.
- [28] LEMBER, J., MATZINGER, H., AND VOLLMER, A. Optimal alignments of longest common subsequences and their path properties. *Bernoulli* 20, 3 (2014), 1292 – 1343.
- [29] MAIER, D. The complexity of some problems on subsequences and supersequences. *J. ACM.* 25, 2 (1987), 322 – 336.
- [30] MALASPINAS, A. S., ERIKSSON, N., AND HUGGINS, P. M. Parametric analysis of alignment and phylogenetic uncertainty. *Bull. Math. Biol.* 73, 4 (2011), 795 – 810.
- [31] MARTIN, J. B. Limiting shape for directed percolation models. *Ann. Probab.* 32, 4 (2004), 2908 – 2937.
- [32] MASEK, W. J., AND PATERSON, M. S. A faster algorithm computing string edit distances. *J. Comput. Syst. Sci.* 20, 1 (1980), 18 – 31.
- [33] MYERS, E. W., AND MILLER, W. Optimal alignments in linear space. *Comput. Appl. Biosci.* 4, 1 (1988), 11–17.
- [34] NEEDLEMAN, S. B., AND WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48 (1970), 443 – 453.
- [35] NG, P. C., AND HENIKOFF, S. Predicting deleterious amino acid substitutions. *Genome Res.* 11, 5 (2001), 863 – 874.

- [36] O'CONNELL, N., AND YOR, M. Brownian analogues of Burke's theorem. *Stoch. Proc. Appl.* 96 (2001), 285 – 304.
- [37] PACHTER, L., AND STURMFELS, B. Parametric inference for biological sequence analysis. *Proc. Natl. Acad. Sci. U.S.A.* 101, 46 (2004), 16138 – 16143.
- [38] PACHTER, L., AND STURMFELS, B. *Algebraic Statistics for Computational Biology*. Cambridge University Press, New York, NY, USA, 2005.
- [39] PRIEZZEV, V. B., AND SCHÜTZ, G. M. Exact solution of the Bernoulli matching model of sequence alignment. *J. Stat. Mech. Theor. Exp.* 2008, 09 (2008), P09007.
- [40] SEPPÄLÄINEN, T. Increasing sequences of independent points on the planar lattice. *Ann. Appl. Probab.* 7, 4 (1997), 886 – 898.
- [41] SEPPÄLÄINEN, T. A scaling limit for queues in series. *Ann. Appl. Probab.* 7, 4 (1997), 855 – 872.
- [42] SMITH, T. F., AND WATERMAN, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* 147 (1981), 195 – 197.
- [43] TRACY, C. A., AND WIDOM, H. Level-spacing distributions and the airy kernel. *Comm. Math. Phys.* 159, 1 (1994), 151 – 174.
- [44] VINGRON, M., AND WATERMAN, M. S. Sequence alignment and penalty choice. review of concepts, case studies and implications. *J. Mol. Biol.* 235, 1 (Jan 1994), 1–12.
- [45] XIA, X. *Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics and Transcriptomics*. Springer, 2007.

NICOS GEORGIU, UNIVERSITY OF SUSSEX, DEPARTMENT OF MATHEMATICS, FALMER CAMPUS, BRIGHTON BN1 9QH, UK.

E-mail address: n.georgiou@sussex.ac.uk

URL: <http://www.sussex.ac.uk/profiles/329373>

JANOSCH ORTMANN, CENTRE DE RECHERCHES MATHÉMATIQUES (CRM), UNIVERSITÉ DE MONTRÉAL, CASE POSTALE 6128, SUCCURSALE CENTRE-VILLE MONTRÉAL (QUÉBEC) H3C 3J7, CANADA.

E-mail address: ortmann@crm.umontreal.ca

URL: <http://crm.umontreal.ca/~ortmann/>